WPPSI®–IV: Equivalence of Q-interactive and Paper Formats

Q-interactive Technical Report 14

Lisa Whipple Drozdick, PhD

Kristen Getz, MA, CCC-SLP

Susan Engi Raiford, PhD

Ou Zhang, PhD

June, 2016

Introduction

This technical report provides information about the adaptation of the Wechsler Preschool and Primary Scale of Intelligence–Fourth Edition (WPPSI–IV; Wechsler, 2012) subtests into digital format for Q-interactive™, Pearson's digital test administration and scoring platform.

Maintaining Format Equivalence

The primary goal when adapting the WPPSI–IV was to maintain raw-score format equivalence between the subtests (i.e., equivalence across paper and digital administration and scoring formats). The effects of digital assessment are minimized through continuing to use external manipulatives and maintaining similar administration and scoring procedures.

When format equivalence is demonstrated, the norms, reliability, and validity information gathered for the paper format of a test can be applied to the digital format. Format equivalence has been demonstrated for the other Wechsler intelligence scales (Daniel, 2012a, 2012b, Daniel, Wahlstrom, & Zhang, 2014; Raiford et al., 2016) and for other tests for young children (Daniel, 2013).

In all of the equivalence studies, it is assumed that digitally assisted (Q-interactive) administration may affect test scores for a number of possible reasons, including the following:

- Examinee interaction with the tablet. To minimize the effects of examinee–tablet interaction that might threaten equivalence, most physical manipulatives are still used with the Q-interactive administration.
- Examiner interaction with the tablet, especially during response capture and scoring. Most of the differences between paper and Q-interactive administrations occur in the examiner interface. Administering a test on Q-interactive is different from a paper administration because Q-interactive includes tools and procedures designed to simplify and support the examiner's task. Great care has been taken to ensure that these adaptations enhance, and do not diminish, the accuracy with which the examiner presents instructions and stimuli, monitors and times performance, and captures and scores responses.
- Global effects of the digital assessment environment. Global effects go beyond just the examinee's or examiner's interaction with the tablet. For example, a global effect was observed in an early study in which the examiner used a keyboard to capture the examinee's verbal responses. Examinees appeared to slow the pace of their responses, or even give shorter responses, so as to avoid having to wait for examiners who were slower typists to finish keying in their verbatim responses.

In the Q-interactive studies, if a task is not equivalent across the two formats, the cause of the format effect is investigated. Understanding the cause is critical to deciding if adjustments by format are required. In principle, if Q-interactive makes examiners more accurate in their administration or scoring, then Q-interactive provides an advance in assessment technology. A reasonable objective for a new technology is for it to produce results equivalent to those from examiners who accurately administer and score the test in paper format. The digital format should not replicate administration or scoring errors that occur in the standard format. If it appears that a digital effect is due to a reduction in accuracy on the part of either the examinee or the examiner, then the first priority is to modify the Q-interactive system to remove this source of error. Only if that were not possible would a normative adjustment be indicated.

In early equivalence studies for Q-interactive, most administrations were video recorded and showed the examiner's and examinee's interactions with their tablets. This provided a way to check the accuracy of administration, recording, and scoring in both digital and paper formats if a format effect (i.e., non-equivalence) was found, so that the cause of the difference could be identified and corrected. These recordings had the additional benefit of showing how examiners and examinees interacted with the test materials in each format, yielding information that could improve interface design. Given the extensive experience that the Q-interactive team acquired over two years and over a thousand administrations, only a portion of the administrations in the present study were video recorded.

As a whole, the equivalence studies indicate that examinees ages 5 and older respond in a similar way when stimuli are presented on a digital tablet rather than in paper components, or when their touch responses are captured by the screen rather than through examiner observation. The present study of the WPPSI–IV provides an opportunity to examine the replicability of this finding among very young children.

Equivalence Study Designs

Several experimental designs have been employed in Q-interactive equivalence studies. In most of them, each examinee takes a subtest only once, in either digital or paper format. This approach avoids the possibility that individual differences in practice effects would impact the results. This method identifies effects that the format may have on how the examinee interacts with the task when they encounter it for the first time. Study designs in which each examinee takes the test only once closely approximate a realistic testing experience.

A number of Q-interactive format equivalence studies have utilized a research design similar to the present study. For example, the Wechsler Adult Intelligence Scale–Fourth Edition (WAIS–IV; Wechsler, 2008), Wechsler Intelligence Scale for Children–Fourth Edition (WISC–IV; Wechsler, 2003), and WISC–V equivalence studies (Daniel, 2012a, 2012b, Daniel et al., 2014, Raiford et al., 2016) relied primarily on an *equivalent-groups* design, with either random or nonrandom assignment of examinees to groups. This design compares the performance of two groups, one taking the test in the digital format and the other in the paper format. The equivalent-groups design is described in detail in Q-interactive Technical Reports 1 and 2. This design, with nonrandom assignment, has been used for the study of the WPPSI–IV.

For all Q-interactive equivalence studies, an effect size of 0.2 or smaller has been used as the standard for equivalence. Effect size is the average amount of difference between scores on Q-interactive and paper administrations, divided by the standard deviation of scores in the population. An effect size of 0.2 is 0.6 of a scaled-score point on the commonly used subtest metric that has a mean of 10 and standard deviation of 3, and is well within the standard error of measurement of the test.

Selection of Participants

The Q-interactive equivalence studies (including this one) use samples of nonclinical examinees. Various studies have examined the performance of special groups of examinees (i.e., intellectually gifted, intellectual disability, specific learning disorders, ADHD, autism spectrum disorder, and motor impairment) on the digital format of the WISC–V (Raiford, Drozdick, & Zhang, 2015; Raiford, Holdnack, Drozdick, & Zhang, 2014; Raiford et al., 2016). These studies provide strong evidence that groups of examinees with these special conditions perform consistently whether tested in paper or digital format.

Examiners participating in the equivalence studies were trained in the tests' standard paper administration procedures. Experience suggests that becoming thoroughly familiar with a new format takes practice. All examiners received enough training and practice in the digital administration and scoring procedures to be able to conduct the administration and capture responses smoothly, without having to devote a great deal of attention to the format.

WPPSI–IV Equivalence Study

Method

Measures

The WPPSI–IV is a comprehensive instrument used to assess intellectual ability for ages 2:6 to 7:7. The adaptation of the WPPSI–IV for digital format was a process that unfolded over a period of three years from February of 2013 to May of 2016, following the publication of the paper format in late 2012. Many of the WPPSI–IV subtests on Q-interactive are nearly identical in administration and scoring to ones in the WISC–V, which has already been evaluated for equivalence (Daniel et al., 2014; Raiford et al., 2016).

Digital administration of some WPPSI–IV subtests involves components external to Q-interactive. For example, all Processing Speed subtests require a paper response booklet; all subtests with components that are manipulated manually by the examinee (i.e., Block Design, Object Assembly, and Zoo Locations) maintain use of those components; and Picture Memory maintains use of a paper stimulus book. For these subtests, the practitioner device (i.e., the tablet used by the practitioner) is used to enter item-level information into Q-interactive so scaled scores can be determined.

Usability studies conducted on an initial version of the WPPSI–IV on Q-interactive in May and June of 2013 suggested that very young children approached the tasks by exploring the tablet's responsiveness to touch (i.e., touch exploration), rather than focusing on the task itself. This was not surprising, because many early childhood apps assume touch exploration and provide few to no instructions to users. Multiple solutions designed to reduce or prevent touch exploration were studied between August of 2013 through December of 2014 (e.g., a constraining tablet case, alteration of other aspects of the user interface that are typical of instruments on Q-interactive that are used with examinees in other age ranges). Experts in early childhood technology interaction were consulted throughout this period. It became clear that touch exploration was not a response to the tablet itself, but was evoked by fascination with and enjoyment of reproducing the classic visual touch-state of Q-interactive stimuli that is seen on so many other Q-interactive tests (i.e., a darkened circle that appears momentarily then disappears to indicate to an examinee that he or she has touched a portion of the stimuli and thereby successfully indicated a choice). The visual touch-state responses were removed from the WPPSI–IV on Q-interactive before the present equivalence study data were collected from January through August, 2015.

Because equivalence had not yet been examined in preschool children, a set of representative subtests was selected to evaluate equivalence. These tests were selected because they involve different types of examinee interactions with the tablet. For example, Block Design and Picture Naming were selected because they involve viewing static stimuli and performing some action, such as using physical components to build a model to match the stimuli or responding verbally to name the stimuli. Receptive Vocabulary, Matrix Reasoning, and Picture Concepts were

selected because they involve the examinee physically touching images on the tablet. Object Assembly was selected because it involves a new type of examiner scoring procedure on the tablet as opposed to the paper Record Form.

Participants

Digital administrations were conducted January through August, 2015. The examinee sample consisted of nonclinical children within the WPPSI–IV age range of 2:6–7:7. Pearson's field research staff recruited examinees and compensated them (i.e., their parents/caregivers) for their participation. Potential participants were screened for demographic characteristics and the same exclusionary factors used to select the WPPSI–IV normative sample (described in chapter 3 of the WPPSI–IV Technical and Interpretive Manual). The sampling plan required approximately equal numbers of males and females, approximately equal numbers of children within each age band (i.e., 2:6–3:11 and 4:0–7:7), and representative proportions of racial/ethnic groups and parent education levels within each of the two age bands.

Potential participants whose characteristics matched the demographic requirements were assigned to the study. A sample of participants that took the WPPSI–IV in paper format was randomly drawn according to the matched demographic variables from the WPPSI–IV normative sample (described in chapter 3 of the WPPSI–IV Technical and Interpretive Manual) to be used for comparison with the digital format group. The paper format group was matched to the digital format group according to age, gender, race/ethnicity, and parent education level. The two samples were used for the analyses.

Examiners were based at numerous locations around the country. All examiners received remote training on the WPPSI–IV and on Q-interactive administration. They conducted practice administrations and received feedback on any administration errors. Examiners who were not Pearson employees were compensated for their participation.

Procedure

Examiners captured response information in the standard manner used for tests on Q-interactive and scored all items (or verified the scores derived by autoscore). The Pearson research team checked cases for completeness and reviewed a portion of the administrations that were captured on video. All subtest total raw scores were calculated automatically by the Q-interactive system.

The format effect on each subtest was estimated by subtracting the digital format group mean scaled score from that of the paper format group. The format effect was converted to an effect size by dividing by the normative-score standard deviation (i.e., 3). The results were used to determine if the format effect for any subtest exceeded the preset criterion of 0.2 or smaller.

Results

Table 1 presents the demographic characteristics of the digital format group. The paper format group's demographic characteristics mirrored those of the digital format group. The group had approximately equal representation of the two WPPSI–IV age bands (i.e., 2:6–3:11 and 4:0–7:7).

Table 1. Demographic Characteristics of the Digital Format Group

N	104
Age	
Mean	4.2
SD	1.0
Range	2:6–6:0
Education	
0–8 years of school	1.0
9–11 years of school, no diploma	10.6
High school diploma or equivalent	11.5
Some college or technical school, associate's degree	33.7
Bachelor's degree	43.3
Gender	
Female	53.8
Male	46.2
Race/Ethnicity	
African American	14.4
Asian	1.9
Hispanic	22.1
Other	6.7
White	54.8

Note. Except for sample size and age, data are reported as percentages. Total percentage may not add up to 100 due to rounding.

Overall, the group had a nearly equal representation of each age band and had nationally-representative proportions of ethnic groups. Females and children of college graduates were somewhat overrepresented relative to the general population. Table 2 reports the means and standard deviations of scores on the selected WPPSI–IV subtests for each format. The magnitude and statistical significance of format effects are also reported. The effect of age was tested and was not found to be a significant contributor to the observed differences, so results from the two age bands were collapsed and reported together.

Table 2. Format Effects: Differences Between WPPSI–IV Scores Obtained Using Paper and Digital Formats

Subtest/ Composite Score	Digital Format		Paper Format						
	Mean	*SD*	Mean	*SD*	*n*	Format Effect	*t* value	*p* value	Effect Size
RV	10.0	3.0	10.4	3.0	104	0.45	1.15	0.25	0.15
PN	11.0	2.9	10.7	2.8	102	-0.31	-0.85	0.40	-0.11
BD	9.6	3.2	10.2	2.9	100	0.60	1.60	0.11	0.20
OA	9.8	3.2	10.0	3.2	95	0.24	0.52	0.61	0.08
MR	10.0	3.5	10.5	2.8	49	0.53	0.87	0.39	0.17
PC	10.1	3.4	10.3	3.4	48	0.23	0.34	0.73	0.07

Subtest abbreviations are: RV = Receptive Vocabulary, PN = Picture Naming, BD = Block Design, OA = Object Assembly, MR = Matrix Reasoning, PC = Picture Concepts.

None of the mean subtest scaled score differences are statistically significant. All effect sizes meet the preset criterion of 0.2 or smaller in absolute value, and therefore, are within the tolerance limits for considering the formats to be equivalent. At the composite score level, the mean differences for the Vocabulary Acquisition Index, Visual Spatial Index, and Fluid Reasoning Index (not shown) are each less than 3 standard score points. These differences are not significant and have negligible effect sizes of 0.2 or smaller in absolute value.

A digital version of the Picture Memory subtest, in which the examinee interacted with the tablet to view stimuli and respond, was administered along with these subtests. In order to minimize construct-irrelevant differences between the paper and digital format, the digital interface on the examinee tablet closely resembled what is seen on WISC–V Picture Span. Because the digital and paper interfaces were so similar, it was assumed that the formats would be equivalent. However, a review of the data and videos of administrations indicated that the correspondence between the paper and digital format was not yet close enough to support the use of digital stimuli and examinee responses. The design work and research related to Picture Memory continued from August of 2015 through the present, but has not yielded results with adequate support. As a result, the initial release of the WPPSI–IV on Q-interactive uses a paper stimulus book and a tablet examiner interface for timing and recording. The examiner interface is similar to the paper format. Because the examinee views a paper stimulus book and the examiner interface is very similar to that of the paper format, Picture Memory was not re-evaluated.

Discussion

Consistent with the overwhelming pattern of results in several previous equivalence studies (e.g., Daniel, 2012a, 2012b; Daniel et al., 2014; Raiford et al., 2016), all format effect sizes fell within the established criterion for Q-interactive format equivalence. This provides evidence that the WPPSI–IV produces consistent scores regardless of format.

Previous research has also shown that there are virtually no statistically significant differences in format effect among subgroups by age, gender, ethnicity, socioeconomic status, or ability level (i.e., Daniel, 2012a, 2012b; Daniel et al., 2014). This set of results indicates that the general finding of an absence of format effects applies broadly to the general nonclinical population.

References

Daniel, M. H. (2012a). *Equivalence of Q-interactive administered cognitive tasks: WAIS–IV* (Q-interactive Technical Report 1). Bloomington, MN: Pearson. Retrieved from http://www.helloq.com/content/dam/ped/ani/us/helloq/media/QinteractiveTechnical%20Report%201_WAIS-IV.pdf

Daniel, M. H. (2012b). *Equivalence of Q-interactive administered cognitive tasks: WISC–IV* (Q-interactive Technical Report 2). Bloomington, MN: Pearson. Retrieved from http://www.helloq.com/content/dam/ped/ani/us/helloq/media/Technical%20Report%202_WISC-IV_Final.pdf

Daniel, M. H. (2013). *Equivalence of Q-interactive and paper administered cognitive tasks: Selected NEPSY–II and CMS subtests* (Q-interactive Technical Report 4). Bloomington, MN: Pearson. Retrieved from http://www.helloq.com//content/dam/ped/ani/us/helloq/media/Technical%20Report%204_NEPSY-II_CMS.pdf

Daniel, M. H., Wahlstrom, D., & Zhang, O. (2014). *Equivalence of Q-interactive and paper administration of cognitive tasks: WISC–V* (Q-interactive Technical Report 7). Bloomington, MN: Pearson.

Raiford, S. E., Drozdick, L. W., & Zhang, O. (2015). *Q-interactive special group studies: The WISC–V and children with autism spectrum disorder and accompanying language impairment or attention-deficit/hyperactivity disorder* (Q-interactive Technical Report 11). Bloomington, MN: Pearson. Retrieved from http://images.pearsonclinical.com/images/assets/WISC-V/Q-i-TR11_WISC-V_ADHDAUTL_FNL.pdf

Raiford, S. E., Holdnack, J. A., Drozdick, L. W., & Zhang, O. (2014). *Q-interactive special group studies: The WISC–V and children with intellectual giftedness and intellectual disability* (Q-interactive Technical Report 9). Bloomington, MN: Pearson. Retrieved from http://www.helloq.com/content/dam/ped/ani/us/helloq/media/Technical_Report_9_WISC-V_Children_with_Intellectual_Giftedness_and_Intellectual_Disability.pdf

Raiford, S. E., Zhang, O., Drozdick, L. W., Getz, K., Wahlstrom, D., Gabel, A., Holdnack, J. A., & Daniel, M. (2016). *WISC–V Coding and Symbol Search in digital format: Reliability, validity, special group studies, and interpretation* (Q-interactive Technical Report 12). Bloomington, MN: Pearson. Retrieved from http://images.pearsonclinical.com/images/Assets/WISC-V/Qi-Processing-Speed-Tech-Report.pdf

Wechsler, D. (2003). *Wechsler intelligence scale for children* (4th ed.). Bloomington, MN: Pearson.

Wechsler, D. (2008). *Wechsler adult intelligence scale* (4th ed.). Bloomington, MN: Pearson.

Wechsler, D. (2012). *Wechsler preschool and primary scale of intelligence* (4th ed.). Bloomington, MN: Pearson.

Wechsler, D. (2014). *Wechsler intelligence scale for children* (5th ed.). Bloomington, MN: Pearson.